Generalized linear models for ordered categorical data

Citation for the original published paper (version of record):
Holm, S. (2023). Generalized linear models for ordered categorical data. Communications in Statistics - Theory and Methods, 52(3): 670-683. http://dx.doi.org/10.1080/03610926.2021.1921210

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
https://doi.org/10.1080/03610926.2021.1921210

Generalized linear models for ordered categorical data

Sture Holm

Department of Mathematical Sciences, Chalmers and Gothenburg University, Gothenburg, Sweden

ABSTRACT

Categorical scale data are only ordinal and defined on a finite set. Continuous scale data are only ordinal and defined on a bounded interval. Due to that character, the statistical methods for scale data ought to be based on orders between outcomes only and not any metric involving distance measure. For simple two-sample scale data, variants of classical rank methods are suitable. For regression type of problems, there are known good generalized linear models for separate categories for a long time. In the present article is suggested a new generalized linear type of model based on non parametric statistics for the whole scale. Asymptotic normality for those statistics is also shown and illustrated. Both fixed and random effects are considered.

ARTICLE HISTORY
Received 6 June 2020
Accepted 19 April 2021

KEYWORDS
Generalized linear model; rank methods; scale data

1. Introduction and summary

The characteristic property of verbal scale data is that it is ordinal on a bounded set. Visual analogue scales are continuous ordinal data bounded to a finite interval. Comparison of two cases can be analyzed truly following the data character by using rank tests. In the classical methods for normally distributed observations, linear models are easily analyzed. For other distributions, the generalized linear models give possibilities to analyze dependence on background variables. For scale data, however, there are not available good simple methods for an overall generalized linear model type of analysis. There exist very good and proper generalized linear model methods for a very detailed analysis for instance of individual levels of the scale and relations between such levels. But they give then separate results for different details.

The present article will suggest a new simple overall type of generalized linear model method, which takes care of the particular scale data type. It is a rank-based method which eliminates the inconvenience of the finiteness of the outcome space by a transformation to an infinite space, in analogy with what is done in a generalized linear model, for example, with a binomial distribution. It will give probabilities to analyze both fixed and random effects.

2. Characteristic properties of scale data

A discrete scale is given by a verbal description of some ordered categories. Here, is an example of a simple pain scale.

CONTACT Sture Holm holm@chalmers.se Department of Mathematical Sciences, Chalmers and Gothenburg University, SE-41296 Gothenburg, Sweden.

1	2	3	4	5	6
No pain	Mild	Moderate	Severe	Very severe	Worst pain possible

The six categories are defined by short verbal descriptions. The categories are also denoted with numbers which they often are in practice. It is not clear, however, that the "distance" between moderate and mild pain should be the same as the "distance" between very severe and severe pain, and should the "distance" between severe pain and moderate pain be "longer" if we introduce a class "rather tough pain" between these two classes.

The common character of ordered category scales is that they have an ordered structure but no real metric structure and it is not easy to find motivation for use of any parametric distribution. I consider the scale data to be a most typical situation for non parametric analysis. The basic statistical model for scale data should include only probabilities for categories and orders between categories, and analysis of the corresponding observations should be based only on relative frequencies of categories and order between them.

Of course, it is convenient to use numbers to denote ordered category data in a computer file but the statistical models and the statistic methods should be invariant under monotone continuous transformation, which is a mathematical description meaning in practice that only the order of the numbers should not have any influence on the models or statistics used in the analysis. There is no standard unit of pain, mobility or quality of life. See also the discussion in Cliff (1996a, 1996b).

The concept invariance is central in the theory of non parametric statistical analysis whose development started some 80 years ago. The intension then was to develop statistical methods for applications where there was no good motivation for the data to be normally distributed or to have any other special parametric distribution. Scale data is a most typical case where there is no good motivations for distribution in any particular parametric class.

Sometimes it is interesting to analyze a simplification of the scale information. For instance, in the pain example above it might be interesting to make a distinction between the scale parts "moderate pain or less" and "severe pain or worse." In a series of independent observations, the number of observations in one of those parts will be binomially distributed and a parametric analysis for that distribution would do well. In such a case a dependence on some background variable(s) could be handled by using an analysis with a generalized linear regression method (McCullagh and Nelder 1989). If you, for instance, have scale data at some time points for individuals it is also possible to include random individual effects in the model when a "cut off method" is used with the scale data.

In most situations with scale observations, there is an interest in making comparisons. It may be a comparison between treatments, comparison of results on different times and so on. We start now with the most basic situation comparing two cases by using sets of independent scale series. Suppose that there are k ordered categories in the scale and use index 1 and 2 for the cases. Then the general model is that observations in series 1 have some category probabilities p_{11}, p_{12}, ..., p_{1k} and the observations in series 2 have some possibly other category probabilities p_{21}, p_{22}, ..., p_{2k}. Since both

series of probability add up to one, there are $2(k-1)$ parameters in this general model. If they were known every probabilistic problem would in principle be possible to solve.

In an application situation, it would be interesting to have fewer parameters and to describe the cases that one of the series has a tendency for higher scale values than the other. And to be true to the data character the method should only depend on the order of observations and not the numbers used to indicate those orders.

If we consider the combination of outcome in category number i for the first case and number j for the second case its probability is $p_{1i}\ p_{2j}$. If $i=j$ the scale category is the same in the two cases, if $i<j$ the second case has a higher category than the first case and if $i>j$ the first case has a higher category than the second case. Adding up the product probabilities for all events where we get the overall probability

$$P_{2>1}=\sum_{i=1}^{k}\sum_{j=1}^{k}p_{1i}\ p_{2j}\ I(i<j)$$

that outcome in case 2 is higher than outcome in case 1, which is a natural alternative if we make a test of no difference between the cases 1 and 2 with the intension to show that case 2 has a distribution with a tendency of higher values than case 1. Here, $I()$ is notation for the indicator of the condition in the parenthesis.

When we want to estimate those parameters to get a test statistic, we can use the relative frequencies in two series of observations for the two cases. If we have n_1 scale observations of the first case and get the random numbers $N_{11},\ N_{12},...,\ N_{1k}$ of observations in the k ordered categories we estimate $p_{1i}\approx\frac{N_{1i}}{n_1}$ for the first series and in the same way $p_{2i}\approx\frac{N_{2i}}{n_2}$ for the second series. Thus, we use the statistic

$$U_2=\sum_{i=1}^{k}\sum_{j=1}^{k}\frac{N_{1i}}{n_1}\ \frac{N_{2j}}{n_2}\ I(i<j)$$

This is a normalized form of the old Mann–Whitney rank test, which uses the inversion sum $\sum_{i=1}^{k}\sum_{j=1}^{k}N_{1i}\ N_{2j}\ I(i<j)$ as a test statistic. A test for an alternative in the other direction uses $U_1=\sum_{i=1}^{k}\sum_{j=1}^{k}\frac{N_{1i}}{n_1}\ \frac{N_{2j}}{n_2}\ I(i>j)$. Alternatively, a two-sided test can be based on the symmetric

$$U=\sum_{i=1}^{k}\sum_{j=1}^{k}\frac{N_{1i}}{n_1}\ \frac{N_{2j}}{n_2}\ sign(i<j)=U_2-U_1$$

where $sign(i<j)=\begin{cases}1\ \ if\ \ i<j\\0\ \ if\ \ i=j\\-1\ \ if\ \ i>j\end{cases}$. Since we work with discrete variables it may be natural to count equal values as "half up and half down" and we will change the above definitions to

$$U_1=\sum_{i=1}^{k}\sum_{j=1}^{k}\frac{N_{1i}}{n_1}\ \frac{N_{2j}}{n_2}\ I(i>j)+\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}\frac{N_{1i}}{n_1}\ \frac{N_{2j}}{n_2}\ I(i=j)$$

and

$$U_2 = \sum_{i=1}^{k}\sum_{j=1}^{k}\frac{N_{1i}}{n_1}\ \frac{N_{2j}}{n_2}\ I(i<j) + \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}\frac{N_{1i}}{n_1}\ \frac{N_{2j}}{n_2}\ I(i=j) = 1 - U_1$$

These mean inversion test statistics may also be described by mean rank statistics. For a possible outcome point, it is defined as the number of strictly smaller observations and half of the number of observations in that outcome. Dividing by the number of observation gives the relative mid-rank. The statistic U_1 can be interpreted as the mean of the relative mid-rank in series 1 corresponding to outcome in series 2. That is the relative mid-ranks in series 1 serves as a "metric" in series 2.

In Svensson and Holm (1994) is defined a parameter relative position RP which is a normalization of $P_{2>1} - P_{1>2}$ making the possible values to be the whole interval $(-1;1)$. The concept is well developed in detail in Svensson (1993). In the present article, I will not make such a normalization, but work with the pure "half-split" of equal observations as described above and work with basic probability parameters $P_{2>1} + \frac{1}{2}P_{1=2}$ and $P_{1>2} + \frac{1}{2}P_{1=2}$.These parameters are estimated by U_2 and U_1. See also Cliff (1996a, 1996b) for discussion of these parameters.

It is convenient for the following to have a notation for the "half-step cumulative distribution" for the scale observations. Define $G_x(i)$ for a value x of the background variable to be the probability of a scale observation to be smaller than category i plus half of the probability of an outcome at category i. These probability parameters are naturally estimated with corresponding relative frequencies. They will be basic quantities in our later discussion of the analysis of scale data in more structured models.

If we have an investigation with several series of independent scale observations, all probabilistic calculations are principally based on the corresponding multinomial distributions. Let$p$ be a vector of true probabilities of all k categories and let $\hat{p}$ the vector of the relative frequencies of the categories in a series of n independent observations.

Then $\hat{p}$ has an asymptotic (k-dimensional) normal distribution with expectation p and covariance matrix C with diagonal elements $\frac{p_i(1-p_i)}{n}$ and off-diagonal elements $-\frac{p_i\ p_j}{n}$. And using the theory of asymptotic normality of regular functions of a basic set of asymptotically normal variables, we can principally find asymptotic distributions of regular functions of scale data by using these general multinomial distributions. See Serfling (2002, subsections 2.7 and 3.3).

3. Generalized linear regression for independent series of scale data

Suppose as an introductory example that we make 10 observations in a scale with five ordered categories at each of four levels of a dose in treatment. In the experimental design, either only 10 patients are used successively at all four dose levels, or 40 patients distributed with 10 patients at each level. We, here, supposed to have the second type of design.

If the observations were regular physical measurements possibly with normal distribution, the analysis could be done with classical regression theory. For other distributions, there is the possibility of using generalized linear regression analysis. For instance, if the observations are in the form of a certain effect appearing it should be a

generalized linear model for binomial distributions. If it is suitable to make a cutoff division of the scale results in two parts, it would be used for scale data results too. Often, however, such a reduction is a waste of information.

There exist several methods of the generalized linear regression type for a detailed analysis of scale data. Many of these methods are given in the book by Agresti (2002) and in Liu and Agresti (2005). Examples of methods are logit models for adjacent responses or scale levels. The methods give a very detailed picture of scale data results principally with different regressions in different parts of the scale. My aim here is to suggest a generalized linear model which can give a picture of the dependence of a few basic non parametric comparison parameters.

As in generalized linear models for parametric families, we want to use a link transformation where the linear model part has no bounds. In the case of continuous data, we may use the log-odds for positive change

$$L(x,\theta) = \ln\left(\frac{P_{2>1}}{P_{1>2}}\right)$$

For scales with discrete classes, it is reasonable to split the probability of equality as above into equal parts joined with increase and decrease and use

$$L(x,\theta) = \ln\left(\frac{P_{2>1}+\frac{1}{2}P_{2=1}}{P_{1>2}+\frac{1}{2}P_{2=1}}\right)$$

where we have the expectations $E[U_1] = P_{1>2} + \frac{1}{2}P_{1=2}$ and $E[U_2] = P_{2>1} + \frac{1}{2}P_{1=2}$. Both the basic parameters $P_{1>2} + \frac{1}{2}P_{1=2}$ and $P_{2>1} + \frac{1}{2}P_{1=2}$ are bound to the interval $[0;1]$ and add to 1. They both may depend on the parameter θ.

As in generalized linear models for parametric families, we want to use a link transformation where the linear model part has no bounds. The same condition applies to the above type of generalized linear model. The Mann and Whitney (1947) test statistic measures only differences between cases. To get a full generalized linear model we need to compare all cases with a suitable basic distribution. This can, for instance, be a placebo, non treatment case, when we have an investigation of the dependence on the dose. Or it can be a joined distribution for all the separate cases appearing in the design.

In the example, we have four series with 10 observations in each, all independent. Then, we can use all 40 observations as the basic comparison distribution. This technique may be compared to the classical Kruskal–Wallis method for continuous distributions (Kruskal 1952; Kruskal and Wallis 1952, 1953), where the mean rank overall series is the comparison element for all different series. This means in practice that a series of observation is compared to the observations in all other series. There is a direct relation between rank differences and inversion sums. For our regression type of analysis, it is, however, most convenient to work with the inversion sums and the joined distribution as a base.

In the following lemma are given some properties which we need to obtain statistical properties of our method.

Lemma 1. *Suppose that* $N_{1,1}$, $N_{1,2}$, ..., $N_{1,k}$ *and* $N_{2,1}$, $N_{2,2}$, ..., $N_{2,k}$ *are the number of outcomes in the ordered categories of two series of with category probabilities* $p_{1,1}$, $p_{1,2}$, ..., $p_{1,k}$ *and* $p_{2,1}$, $p_{2,2}$, ..., $p_{2,k}$ *for* $n_1 = N_{1,1} + N_{1,2} + \ldots + N_{1,k}$ *and*

$n_2 = N_{2,1} + N_{2,2} + ... + N_{2,k}$ scale observations in the two series and that all scale observations are independent. Further, let $\hat{p}_1$ and $\hat{G}_2$be vectors with components

$$\hat{p}_{1i} = \frac{N_{1i}}{n_1}$$

and

$$\hat{G}_2(i) = \frac{\sum_{j=1}^{i-1} N_{2,j} + \frac{1}{2} N_{2,i}}{n_2}$$

Then, the inversion means $U_1 = \hat{p}_1^T \hat{G}_2$ and $U_2 = 1 - U_1$ are asymptotic normally distributed as n_1 and n_2 tend to infinity. The asymptotic variance for U_1 and U_2 are both estimated by

$$\hat{G}_2^T C_1 \hat{G}_2 + \hat{G}_1^T C_2 \hat{G}_1$$

where C_1 and C_2 are the asymptotic covariance matrices for $\hat{p}_1$ and $\hat{p}_2$.

Proof. The proof is based on Serfling (2002, subsections 2.7 and 3.3). Using differentials and the fact that $U_1 = \hat{p}_1^T \hat{G}_2 = 1 - \hat{p}_2^T \hat{G}_1$ we see that the asymptotic distribution has the same asymptotic variance as the one of

$$\left(\hat{p}_1^T - p_1^T\right) G_2 - \left(\hat{p}_2^T - p_2^T\right) G_1$$

These parts are independent and by basic multivariate properties, the two variances are

$$\hat{G}_2^T C_1 \hat{G}_2 \text{ and } \hat{G}_1^T C_2 \hat{G}_1.$$

□

Lemma 2. *Suppose that $U_{1,2}$ and $U_{1,3}$ are defined by*

$$U_{1,2} = \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{N_{1i}}{n_1} \frac{N_{2j}}{n_2} I(i > j) + \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{N_{1i}}{n_1} \frac{N_{2j}}{n_2} I(i = j)$$

and

$$U_{1,3} = \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{N_{1i}}{n_1} \frac{N_{3j}}{n_3} I(i > j) + \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{N_{1i}}{n_1} \frac{N_{3j}}{n_3} I(i = j)$$

for three independent scale series. In vector notations $U_{1,2} = \hat{p}_1 \hat{G}_2$ and $U_{1,3} = \hat{p}_1 \hat{G}_3$.

Their asymptotic covariance equals $G_2^T C_1 G_3$.

Proof. With the same technique as in Lemma 1 we can see that the required asymptotic covariance is the same as the asymptotic covariance between $(\hat{p}_1^T - p_1^T)G_2 - (\hat{p}_2^T - p_2^T)G_1$ and $(\hat{p}_1^T - p_1^T)G_3 - (\hat{p}_3^T - p_3^T)G_1$. Since the three random vectors $\hat{p}_1^T - p_1^T$, $\hat{p}_2^T - p_2^T$ and $\hat{p}_3^T - p_3^T$ are independent, that asymptotic covariance equals the asymptotic covariance between $(\hat{p}_1^T - p_1^T)G_2$ and $(\hat{p}_1^T - p_1^T)G_3$, which is equal to $G_2^T C_1 G_3$. □

Now, consider a case where we make independent scale observations for a number K of values of some background variable(s). The above Lemmas will, of course, hold with the indexes 1 and 2 changed to any index in the set 1, 2, ..., K.

When we consider the test statistic for any comparison case μ with the alignment of all cases we start by writing the cumulative distribution function G_0 of the totality, which is a vector with components

$$G_{0i} = \frac{1}{n_0}\sum_{j=1}^{m} n_j G_{ji}$$

for all $i = 1, 2, ..., k$. Then, the asymptotic of the combined $U_{\mu 0}$ statistic $U_{\mu 0} = \hat{p}_\mu^T \hat{G}_0 = \hat{p}_\mu^T \sum_{j=1}^{m} \frac{n_j}{n} \hat{G}_j$ for a case μ is the same as the asymptotic variance for

$$\begin{aligned}
&\left(\hat{p}_\mu - p_\mu\right)^T G_0 - \frac{1}{n_0}\sum_{j=1}^{m} n_j \left(\hat{p}_j - p_j\right)^T G_\mu \\
&= \left(\hat{p}_\mu - p_\mu\right)^T \left(\frac{1}{n_0}\sum_{j=1}^{m} n_j G_j\right) - \frac{1}{n_0}\sum_{j=1}^{m} n_j \left(\hat{p}_j - p_j\right)^T G_\mu \\
&= \left(\hat{p}_\mu - p_\mu\right)^T \left(G_0 - \frac{n_1}{n_0} G_\mu\right) - \frac{1}{n_0}\sum_{j=1, j\neq\mu}^{m} n_j \left(\hat{p}_j - p_j\right)^T G_\mu
\end{aligned}$$

We calculate its asymptotic variance taking into consideration the rules for multivariate statistics. Then, for any index μ

$$\begin{aligned}
&\mathrm{Var}\left(\left(\hat{p}_\mu - p_\mu\right)^T G_0 - \frac{1}{n_0}\sum_{j=1}^{m} n_j \left(\hat{p}_j - p_j\right)^T G_\mu\right) \\
&= \mathrm{Var}\left(\left(\hat{p}_\mu - p_\mu\right)^T \left(G_0 - \frac{n_\mu}{n_0} G_\mu\right)\right) + \frac{1}{n_0^2}\sum_{j=1, j\neq\mu}^{m} n_j^2 \mathrm{Var}\left(\left(\hat{p}_j - p_j\right)^T G_\mu\right) \\
&= \left(G_0 - \frac{n_\mu}{n_0} G_\mu\right)^T C_\mu \left(G_0 - \frac{n_\mu}{n_0} G_\mu\right) + \frac{1}{n_0^2}\sum_{j=1, j\neq\mu}^{m} n_j^2 G_\mu^T C_j G_\mu
\end{aligned}$$

If μ and ν are different indices the asymptotic covariance $Cov(\hat{p}_\mu^T G_0, \hat{p}_\nu^T G_0)$ equals

$$\begin{aligned}
&Cov\left(\left(\hat{p}_\mu^T - p_\mu^T\right)\left(G_0 - \frac{n_\mu}{n} G_\mu\right) - \frac{n_\nu}{n_0}\left(\hat{p}_\nu^T - p_\nu^T\right) G_\mu, \left(\hat{p}_\nu^T - p_\nu^T\right)\left(G_0 - \frac{n_\nu}{n} G_\nu\right)\right. \\
&\left. - \frac{n_\mu}{n_0}\left(\hat{p}_\mu^T - p_\mu^T\right) G_\nu\right) \\
&= \frac{n_\nu}{n_0}\left(G_0 - \frac{n_\mu}{n} G_\mu\right)^T C_\mu G_\nu + \frac{n_\mu}{n_0}\left(G_0 - \frac{n_\nu}{n} G_\nu\right)^T C_\nu G_\mu
\end{aligned}$$

Finally, we have the log-odds transformation of the variables $U_{j0} = \hat{p}_j{}^T \hat{G}_0$. The derivative of the function $f(u) = \ln\left(\frac{u}{1-u}\right)$ is $f'(u) = \frac{1}{u} + \frac{1}{1-u} = \frac{1}{u(1-u)}$, and the estimate of the scale factor in the transformation of the asymptotic distribution is just this derivative in the observation point for the different variables $U_{j0} = \hat{p}_j{}^T \hat{G}_0$, we may now formulate

the final result. If we have calculated the estimated variance-covariance matrix Σ_0 for the variables $U_{j0} = \hat{p}_j{}^T \hat{G}_0$ we get the estimated variance-covariance matrix Σ_1 for the transformed variables by

$$\Sigma_1 = \Lambda \ \Sigma_0 \ \Lambda$$

where Λ is a diagonal matrix with diagonal elements $\frac{1}{U_{j0}(1-U_{j0})}$ for $j = 1, \ 2, \ ..., \ m$.

We now have all tools for making a general linear statistical analysis. The asymptotic variance of any linear function of the basic logodds-values $\log(U_{j0}/(1-U_{j0}))$ can be determined from Σ_1 by simple matrix calculations. Let the coefficients of the linear function be used as components in the column vector L. Then, by elementary rules in multivariate statistics the asymptotic variance for the estimate equals $\Sigma_2 = L^T \ \Sigma_1 \ L$.

In the statistical analysis, the important test of the hypotesis that a parameter is 0 can be based on the ratio between the parameter and its estimated standard error $\sqrt{L^T \ \Sigma_1 \ L}$ which has an approximate normal (0,1) distribution if the hypothesis is true. Approximate confidence interval for a parameter is easily constructed using the same estimates.

If we have an application where there are linear estimates of $K \geq 2$ parameters they all have their own coefficients in the linear estimates. Now let L be a matrix with K columns equal to the coefficient vectors for different parameters. Then direct calculations now shows that $\Sigma_2 = L^T \ \Sigma_1 \ L$ is a matrix having variance estimates of parameter estimated in the diagonal and covariance estimates in the off-diagonal elements. Any multiparameter generalized linear problem can be handled at the price of more complexity. This includes also multiple statistical tests and multiple confidence intervals.

4. Two numerical examples in one

I want now to make two numerical examples, one regression example and one analysis of variance example. To illustrate both types simple, I will use the same four artificially generated series of scale data for both. Many of the calculations are the same anyway.

Suppose that the scale has 7 categories and that we make 25 observations in each series. Here, are the generated data of numbers of observations in categories.

	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6	Cat. 7
Series 1	2	4	7	5	4	2	1
Series 2	0	2	5	7	6	3	2
Series 3	0	0	3	5	9	5	3
Series 4	0	0	0	2	6	9	8
Totally	2	6	15	19	25	19	14

The U statistics for separate series compared to total are easily calculated and the results are:

Series	1	2	3	4
U statistic	0.3028	0.4230	0.5444	0.7298

Observe that the mean of the U statistics is the neutral value 0.5000. Log-odds transformation of the U statistics give

Series	1	2	3	4
Logodds	−0.8340	−0.3105	0.1781	0.9936

Running through the calculations described above, we get the estimate

$$\Sigma_1 = \begin{bmatrix} 0.0401 & 0.0188 & 0.0156 & 0.0118 \\ 0.0188 & 0.0317 & 0.0156 & 0.0131 \\ 0.0156 & 0.0156 & 0.0264 & 0.0139 \\ 0.0118 & 0.0131 & 0.0139 & 0.0259 \end{bmatrix}$$

of the variance-covariance matrix for these estimates. That is the basis for the calculation of variances of parameters in our linear model. However, the treatment of those result now depends on the type of generalized linear model we have.

If our data comes from a situation where our statistical model is a simple linear regression with the background variable equal to the series number, the estimate of the regression function of the log-odds data is

$$\mu(t) \approx 0.0068 + 0.597(t - 2.5)$$

The formulas in the previous section show that the estimate of the asymptotic variance of the steepness estimate is 0.00434 and the corresponding standard error 0.0659. The steepness is significally greater than 0, since it has a test statistic $\frac{0.597}{0.0659} = 9.06$ and a p value very close to 0. So a confidence interval with approximate confidence degree of 95% for the steepness would be

$$0.597 \mp 1.96 \cdot 0.0659 = [0.468; 0.726].$$

When we have a one-dimensional dependence of scale results on some background variable x, it is natural to suppose that there is a stochastic ordering of the results determined by the order of the x values. A result Y_2 is said to be stochastically larger than a result Y_1 if their cumulative distribution functions $G_2(y)$ and $G_1(y)$ satisfy

$$G_2(y) \leq G_1(y) \text{ for all and } G_2(y) < G_1(y) \text{ for some } y.$$

It means that the distribution of Y_2 lies more to the right than the distribution of Y_1. Here, is a Figure 1 of the empirical cumulative distribution functions in the example.

In this example, the empirical distributions are stochastically ordered. You should observe, however, that even if the theoretical distributions are stochastically ordered, the empirical can lack that property due to natural randomness, in particular for small sample sizes. More about stochastic ordering is found in Lehmann and Romano (2005).

The Mann–Whitney parameter and empirical statistic are suitable measures of the differences between ordered distributions, but it is more general and works well also with distributions, which are not stochastically ordered. So in our analysis, we do not have to assume that we have stochastic ordering between our distributions. Our method works well anyway.

If our data comes from a two-factor ANOVA design with factor A low in series 1 and 2 and high at series 3 and 4, and factor B low in series 1 and 3 and high in series 2

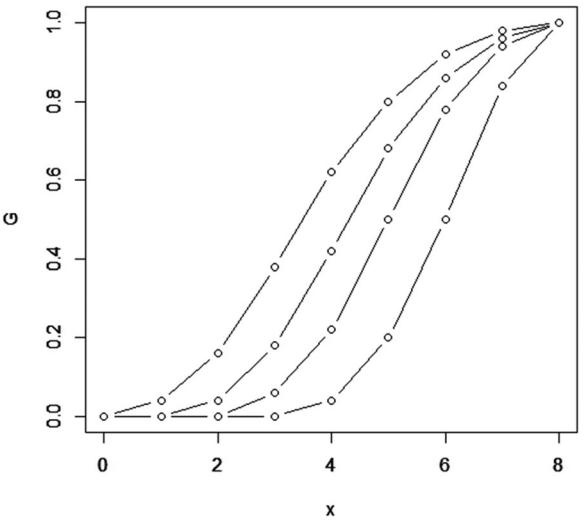

Figure 1. Empirical cumulative distribution functions for the four series in the example. They are seen to be stochastically increasing in the background variable *x*.

and 4, the estimates would be: General mean 0.0068, A effect 0.579 and B effect 0.335. Their estimated standard errors are 0.0696 and 0.0619. Both effects are significantly different from 0 with extremely small p values. If we make simultaneous confidence intervals with multiple confidence degree 95% by the simple Bonferroni method (individual confidence degree 97.5%) we get $[0.423\ ; 0.735]$ and $[0.199\ ; 0.474]$.

The specific scale data calculations are the same for both types of models and further discussion of the linear model are made differently. The calculations needed to get the basic properties for the in-data to the linear model may seem complicated, but using vectors and matrices they are quite straightforward after all and they need no use of any iterative technique.

5. Hierarchies and random effects

The randomness of scale results may often have a hierarchic character. Think for instance of a pedagogic investigation where the student results are given as an ordered scale category. It is of course influenced by the student random error, choice of participants, daily condition and so on. But there is certainly also a teacher effect on the results. And that is common for a whole class. If we want to compare two methods to learn a subject and use a random half of the available classes use method A and the other classes use method B we would need to take into consideration the two levels of randomness in our analysis of the results. And for the reliability of the result of the analysis, it is very important to do so.

In other fields there are similar situations, for example, medical investigations may have a clinic effect beside the individual effect, a technical investigation may have a material batch effect beside the production effect and so on.

To be concrete in the discussion, let us think of a pedagogic investigation with students in classes and comparison between two methods of education for some part of

the curriculum. Let us first consider the class effect to be a fixed background effect. Then, we have here a hierarchy of the method effect and class effect. Technically we can estimate the joined effect of the education method and the class by making an ordinary analysis for the classes. Could be done for all classes for both education methods.

How about the education effects themselves, separated from the class effects? They can easily be handled. Suppose for instance that when we compare two pedagogic arrangements A and B, use 6 classes randomly distributed with 3 for each arrangement, and suppose there are 25 students in each class. In our generalized linear model, we work with a log-odds transformation. In analogy with classical normal linear models and generalized linear models for other distributions, we can suppose that the class effect and the student effect are additive in this scale. Thus, in the present example, we have three transformed class test values for each arrangement. Their respective means estimate the total arrangement effects and the ordinary standard deviations of the class estimates can be used to get a standard errors. It further enables the estimate of arrangement effect difference and test of the hypothesis of equal effect.

Have we now dropped the earlier types of error estimates? No, we have not. We have 6 variance estimates in total, one for each class. They all estimate the variance for the student results variation within their class. If we want to separate the class effect and student effect their mean has to be used. But for comparing the non random arrangement effects we need just to do the simple calculation in the previous paragraph.

It is quite common that the high-level random effects, like the class effect in our example are the dominating ones in an investigation.

6. Continuous ordered scales

In many application fields, are used visual analogue scales, VAS. For instance in pain judgments the estimate of the grade is given by a mark on a line between "Absence of pain" at the left end and "Worst pain experienced" at the right end. These marks are then read off as a number by a millimeter ruler.

The previously discussed methods work equally well in this case as in the case with discrete ordered classes. It may well be that distribution here is of mixed continuous-discrete type since for instance the two endpoints of the scale can work as distinct points. It does not make any difference what concerns the analysis. And millimeter results are just discrete anyway.

It is very important, however, to note that the VAS scale results are only ordered. There is no full metric with a distance. But no harm in that, the previously discussed methods based on ranks work perfectly well and they are also efficient. I will not prove any results for this case here. The proofs will quite easily follow along the same lines as the one for category data. The finiteness of the outcome space makes standard conditions for asymptotic normality be satisfied.

7. Other methods and efficiency questions

Quite often is seen that investigations with scale observations are analyzed by old traditional methods, which are merely based on assumptions of normality, with the same

variance in different series and effects appearing as translations. None of these basic assumptions is reasonable for scale data series. Sums and means and empirical variances may for instance generate confidence intervals including values, which are outside the parameter space.

If we had some well-motivated class of distributions on the set of possible scale results we could preferably make an efficient parametric analysis with this distribution. That would be very good. However, I do not know any successful examples of this type.

I cannot find any application motivation for using the binomial distribution for scale data. Let us anyway consider as an example a scale with five categories and the class of distributions for $0 < p < 1$ which assigns the five binomial $(4,p)$ densities $\binom{4}{k-1}p^{k-1}(1-p)^{4+1-k}$ for $k = 1,\ 2,\ 3,\ 4,\ 5$ to the classes. In a data set the adjustment $\frac{\bar{y}-1}{4}$ of the mean $\bar{y}$ of category numbers would be a good estimate of the parameter p. Can our non parametric method compete with an analysis based on that parametric estimate? To find out we compare the asymptotic power functions, e.g., in the point $p = 0.6$ following the path $p = 0.6 + \Delta$ by using the neighboring points $p = 0.6$ and $p = 0.6 + 0.01$, same sample size n in both cases. The calculations are quite easily done with earlier given formulas. For the parametric method, we get the asymptotic function for a two-sided 5% test

$$\beta_{\text{par}}(\Delta) = \Phi\big(-1.96 + 0.0289\ \Delta\sqrt{n}\big),$$

and for the non parametric method, it is

$$\beta_{\text{non}}(\Delta) = \Phi\big(-1.96 + 0.0286\ \Delta\sqrt{n}\big).$$

The coefficients for Δ differ 1% in favor of the parametric method. Not surprising since the parametric method is fitted exactly to the studied power trace. But what happens if the real data does not fit the parametric model well? As an example, we calculate the power in the direction determined by the two close alternative distributions 0.16, 0.16, 0.16, 0.16, 0.36 and 0.17, 0.17, 0.17, 0.17, 0.32 to find the asymptotic power for the two methods.

The result here is

$$\beta_{\text{non}}(\Delta) = \Phi\big(-1.96 + 0.0505\ \Delta\sqrt{n}\big)$$

for the non parametric method and

$$\beta_{\text{par}}(\Delta) = \Phi\big(-1.96 + 0.0474\ \Delta\sqrt{n}\big),$$

for the parametric method. The coefficient is in favor of the non parametric method with 6.5%. Since the size factor is a square root of n it means that the parametric method needs about 13% more observations than the non parametric method to get the same power. Somewhat bigger difference than in the previous case.

These small calculations show that there may exist cases where a parametric method can have more power than the non parametric method as well as cases where the non parametric method has more power than a parametric method. To give a broader view of the power differences between the inversion sum test and the test based on means of category numbers requires a lot of space for the presentation and it is even hard to find

suitable comparison alternatives. At least the above example indicates that an inversion sum test may have almost the same power as a parametric test if the data fits the parametric model well and the inversion sum test may have quite a bit better power if the data fits the parametric model bad. And I have presented earlier in the article how a generalized linear model can be adapted to the non parametric method. For an application where one can find a trustable one-parameter class of distributions supposed to fit the data well, a parametric method is certainly a good alternative. But a fair comparison would also require a suitable construction of a linking to a linear space for the parametric method, which we have for the non parametric generalized linear model.

The generalized linear models for separate levels of scale data are of cause very good types of analyses, which enables additive analysis. (See e.g., Agresti 2002, chapter 8). However, it involves several different models so it is not directly useful for our problem where we aim at a method which should use one common positional measure for each basic series. The spirit of the present method and those methods are very much the same. One strong reason for our formulation is the ability to handle for instance the influence of random clinic level effects in multicenter investigations, which should influence each scale data series as one unit. And also in the non random part of the generalized linear model, the basic observation series should be treated as one unit describing the position in the calculation of the parameter estimates and statistical properties.

Our method is much related to the Kruskal–Walley rank test of equality of several distributions, which has the test statistic $\sum_k (\bar{R}_k - \bar{\bar{R}})^2$. Discrete distributions require the ranks to be mid-ranks since there are lots of coinciding values. Here, now the difference $\bar{R}_k - \bar{\bar{R}}$ is exactly the same as the mean inversion (Mann–Whitney) test statistic of the case k with the full set of all observations. The rank differences are built up by inversions where coinciding values are counted as half inversions.

Our method includes the well-known and much used Kruskal–Wallis test as a special case. That test itself can just answer the question IF there are any effects. To find out WHAT effects there are in a statistically strict way we need something more. The methods in the present article is aimed at filling that gap in statistical theory by presenting a general method and prove the asymptotic distribution properties needed for using it in practice for example for finding rejection points in *post hoc* tests, for creating simultaneous confidence intervals, make power calculations and plan the sample sizes in investigations with scale observations.

Acknowledgments

I am grateful to an anonymous referee for valuable suggestions.

References

Agresti, A. 2002. *Categorical data analysis*. 2nd ed. New York: Wiley.

Cliff, N. 1996a. Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research* 31 (3):331–50. doi:10.1207/s15327906mbr3103_4.

Cliff, N. 1996b. *Ordinal methods for behavioral data analysis*. Mahwah, NJ: Lawrence Erlbaum.

Kruskal, W. H. 1952. A nonparametric test for the several sample problem. *The Annals of Mathematical Statistics* 23 (4):525–40. doi:10.1214/aoms/1177729332.

Kruskal, W. H., and W. A. Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 47 (260):583–612. doi:10.2307/2281082.
Kruskal, W. H., and W. A. Wallis. 1953. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 48:907–11.
Lehmann, E. L., and J. P. Romano. 2005. *Testing statistical hypotheses*. New York: Springer.
Liu, I., and A. Agresti. 2005. The analysis of ordered categorical data: An overview and a survey of recent development. *Test* 14 (1):1–73. doi:10.1007/BF02595397.
Mann, H. B., and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* 18 (1):50–60. doi:10.1214/aoms/1177730491.
McCullagh, P., and J. A. Nelder. 1989. *Generalized linear models*. 2nd ed. London: Chapman and Hall.
Serfling, R. J. 2002. *Approximation theorems of mathematical statistics*. New York: Wiley.
Svensson, E. 1993. *Analysis of systematic and random differences between paired ordinal categorical data*. Göteborg: Almqvist and Wiksell.
Svensson, E., and S. Holm. 1994. Separation of systematic and random differences in ordinal rating scales. *Statistics in Medicine* 13 (23–24):2437–53. doi:10.1002/sim.4780132308.